What Makes A Successful Grid Application?

Jim Gray (Microsoft)
Alex Szalay (Johns Hopkins)

Presentation at
Earth System Science & Applications Advisory Committee
(NASA - ESSAAC)
San Diego, Ca
18 Feb 2004

Thesis

- Definition:
  - Success == Wide use
- Thesis:
  - Successful Grid App == Content + Applications

Examples

Content	Applications
• Web	• Mosaic, Navigator… AltaVista, Yahoo!,Google ...
• eMail	• AOL, HotMail, …
• BB/Chat	• AOL, ICQ, IM,
• NLM	• Genbank, PubMed, BLAST, Entrez, …
• USGS	• TerraServer, National Map Map Point, Map Quest...
• NASA	• EOS/DIS?

- Download (FTP and GREP) are not adequate
  - You can GREP 1 MB in a second
  - You can GREP 1 GB in a minute
  - You can GREP 1 TB in 2 days
  - You can GREP 1 PB in 3 years.
- Oh!, and 1PB ~10,000 disks
- At some point we need
  indices to limit search
  parallel data search and analysis
- This is where databases can help
- Next generation technique: **Data Exploration**
  - Bring the analysis to the data!

Near-line Data Is Dead Data
Tape Archives = Data Roach Motels

- Tape-based data has bad data access
- Now that disk data is 1M$ / Petabyte
  Tape is not worth the trouble
  More expensive overall

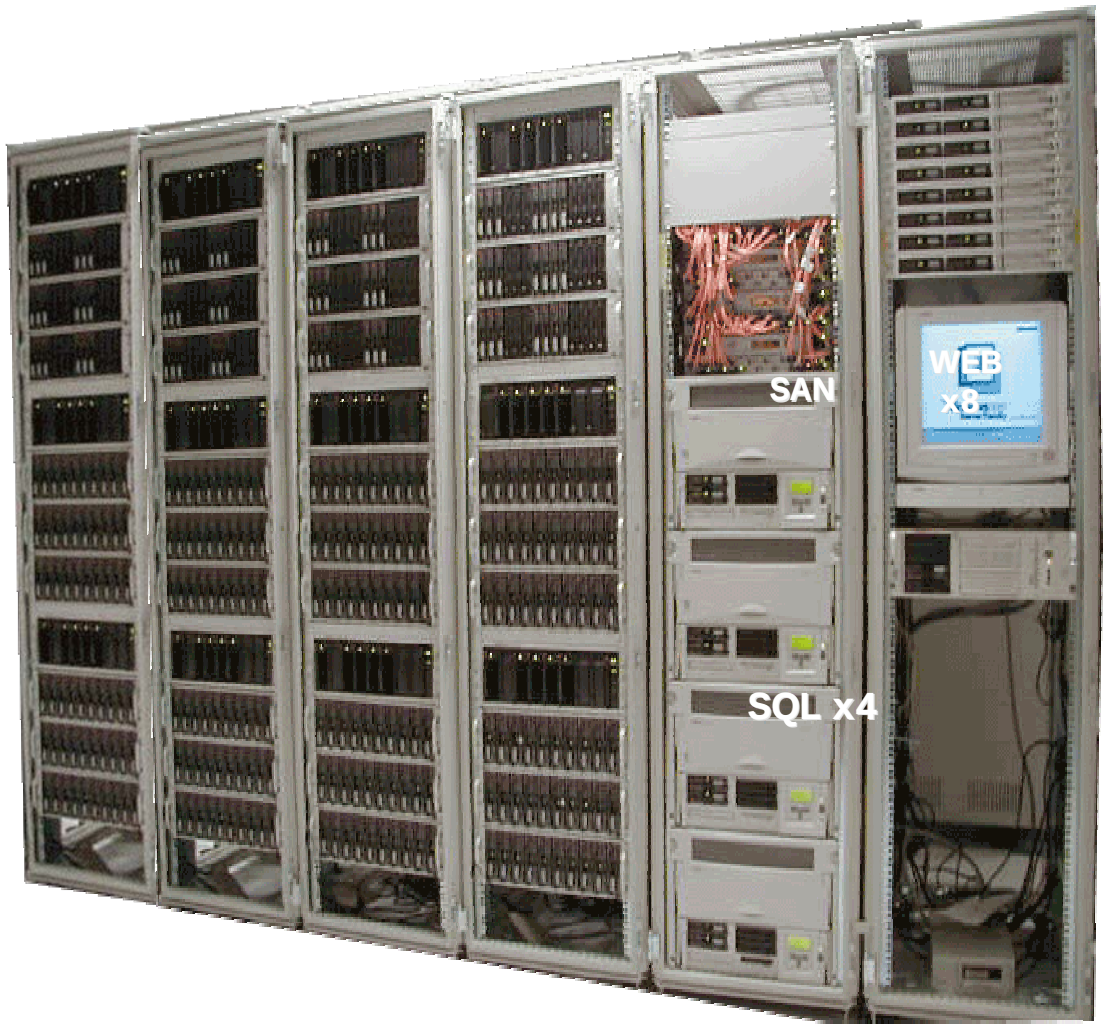
- Need CONTENT + ACCESS

TerraServer: TerraService.Net

- Popular geo-spatial app
  - Averages 40k visitors / day; 1 million map views
  - .NET Web Service OpenGIS web map server
  - Dynamic Map Re-projection
    - UTM to Geographic projection
    - Dynamic texture mapping
- Used in production applications
  - USDA, EPA, FEMA
- New Data
  - 1 foot resolution natural color imagery
  - Census Tiger data
- Lights Out Management
  - MOM
  - Auto-backup / restore on drive failure

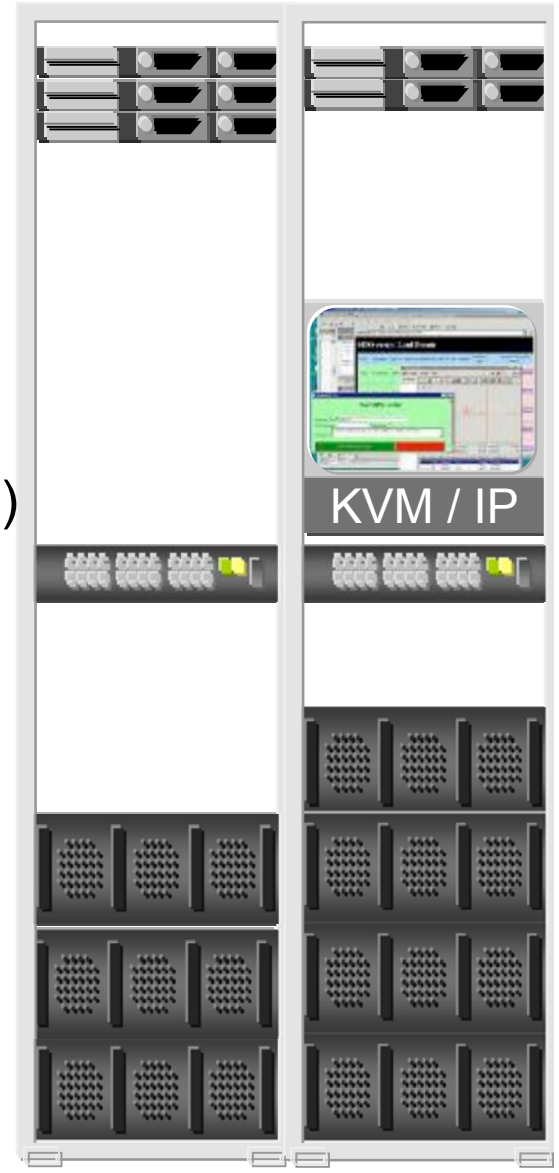
TerraServer V4

- 8 web front end
- 4x8cpu+4GB DB
- 18TB triplicate disks
  Classic SAN
  (tape not shown)
- ~2M$
- Worked GREAT!
- 2000…2003
- Now replaced by V5

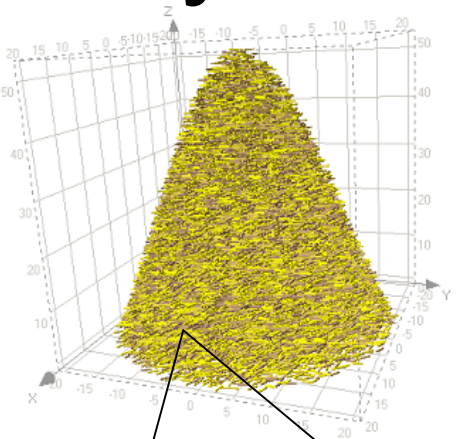


TerraServer V5

- Storage Bricks
  - "White-box commodity servers"
  - 4tb raw / 2TB Raid1 SATA storage
  - Dual Xeon 2.4ghz, 4GB RAM
- Partitioned Databases (PACS – partitioned array)
  - 3 Storage Bricks = 1 TerraServer data ++
  - Data partitioned across 20 databases (~containers)
  - More data & partitions coming
- Low Cost Availability
  - 4 copies of the data
    - RAID1 SATA Mirroring
    - 2 redundant "Bunches"
  - Spare brick to repair failed brick
    2N+1 design
  - Web Application "bunch aware"
    - Load balances between redundant databases
    - Fails over to surviving database on failure
- ~100K$ capital expense.



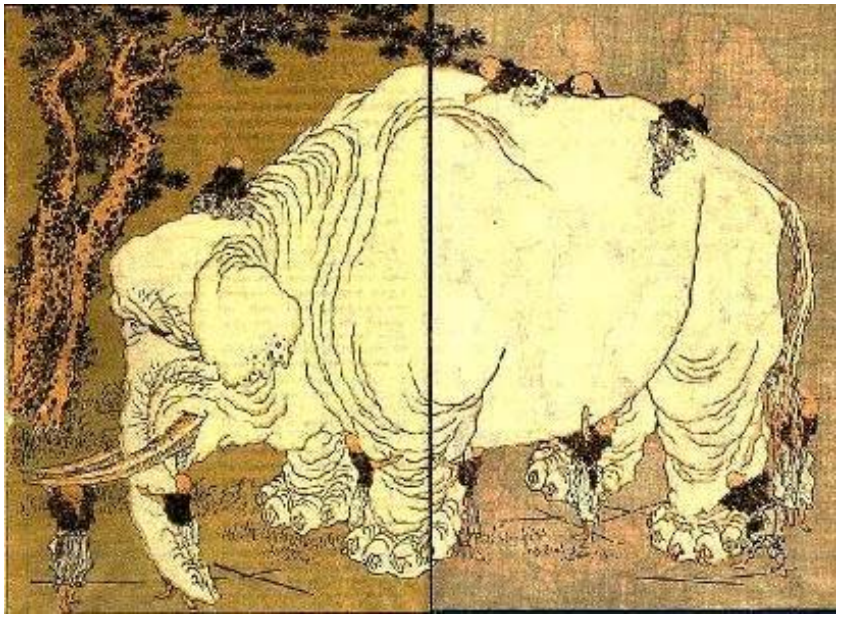
Next-Generation Data Analysis



- Looking for
  - Needles in haystacks – the Higgs particle
  - Haystacks: Dark matter, Dark energy
- Needles are easier than haystacks
- Global statistics have poor scaling
  - Correlation functions are N^2, likelihood techniques N^3
- As data and computers grow at same rate, we can only keep up with $N\,logN$
- A way out?
  - Relax notion of optimal
    (data is fuzzy, answers are approximate)
  - Don't assume infinite computational resources or memory
- Combination of statistics & computer science

Analysis and Databases

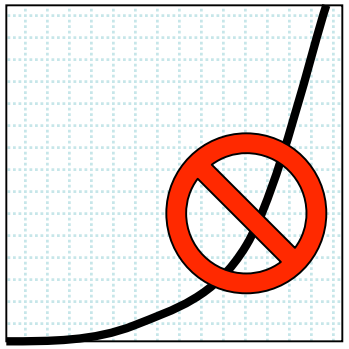
- Much statistical analysis deals with
  - Creating uniform samples –
  - data filtering
  - Assembling relevant subsets
  - Estimating completeness
  - censoring bad data
  - Counting and building histograms
  - Generating Monte-Carlo subsets
  - Likelihood calculations
  - Hypothesis testing



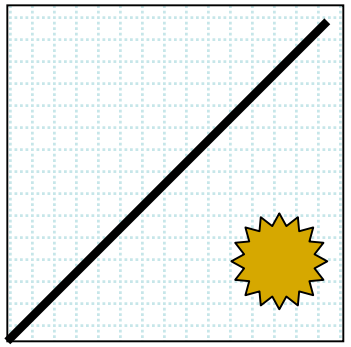
- Traditionally these are performed on files
- Most of these tasks are much better done inside a database
- Move Mohamed to the mountain, not the mountain to Mohamed.

Organization & Algorithms

- Use of clever data structures (trees, cubes):
  - Up-front creation cost, but only *N logN access cost*
  - Large speedup during the analysis
  - Tree-codes for correlations (A. Moore et al 2001)
  - Data Cubes for OLAP (all vendors)



- Fast, approximate heuristic algorithms
  - No need to be more accurate than cosmic variance
  - Fast CMB analysis by Szapudi et al (2001)
    - *N logN* instead of N^3 => 1 day instead of 10 million years



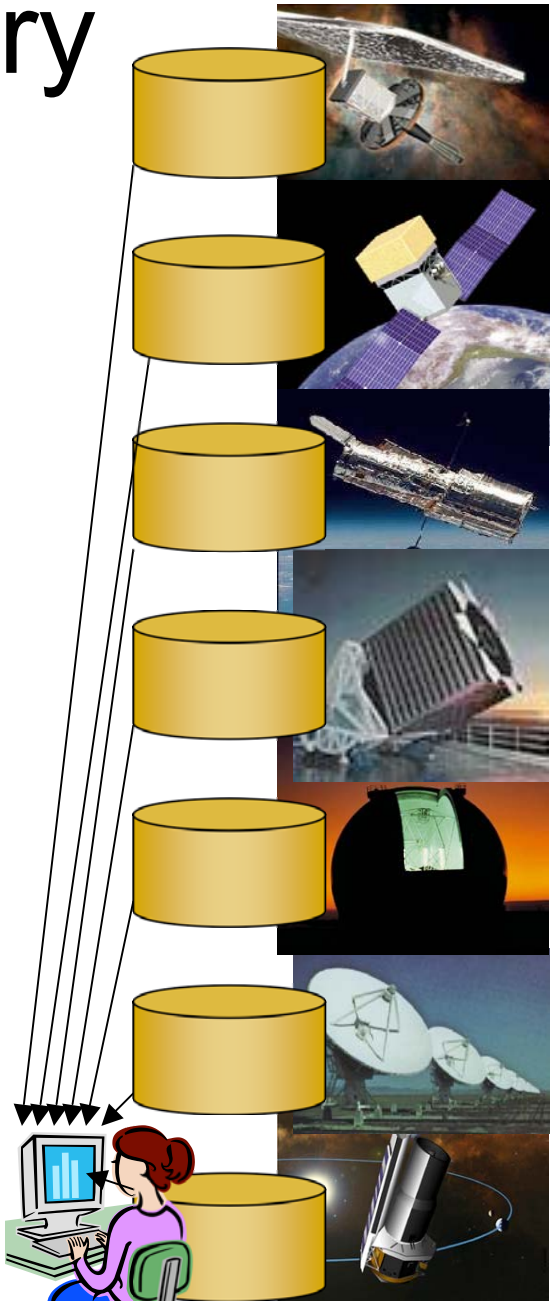
- Take cost of computation into account
  - Controlled level of accuracy
  - Best result in a given time, given our computing resources

Why Is Astronomy Special?

- Data has no commercial value
  - No privacy concerns, freely share results with others
  - Great for experimenting with algorithms
- Data is real and well documented
  - High-dimensional (with confidence intervals)
  - Spatial, temporal
- Diverse and distributed
  - Many different instruments from many different places and many different times
- The questions are interesting
- There is a lot of it (soon Petabytes)

The Virtual Observatory

- Many new surveys are coming
  - SDSS is a dry run for the next ones
  - LSST will be 5TB/night
- All the data will be on the Internet
  - ftp, web services…
- Data and applications will be associated with the projects
  - Distributed world wide, cross-indexed
  - Federation is a must
- The world's best telescope
  - World Wide Telescope
- Finds the "needle in the haystack"
- Successful demonstrations in Jan'03

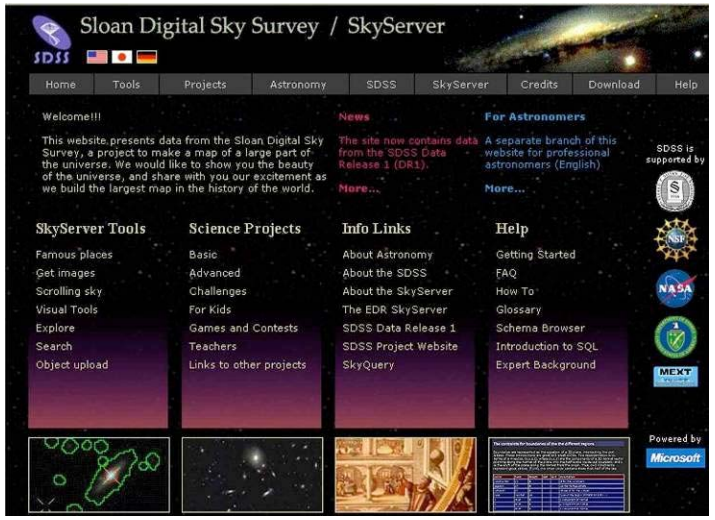


Short History of the VO

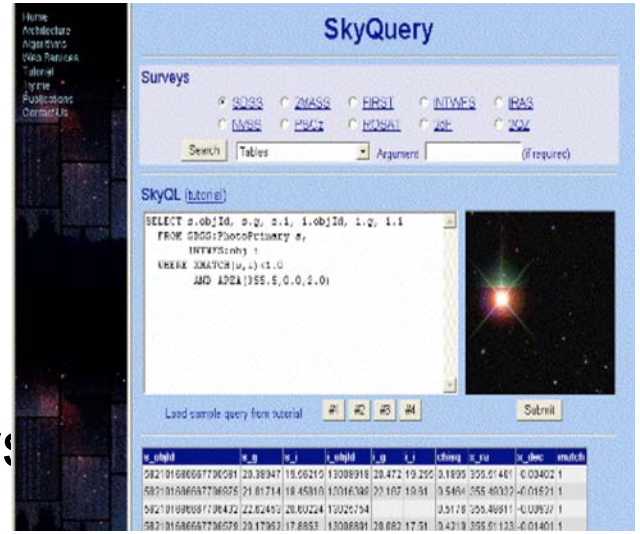
- Driven by exponential data growth
- Started with SDSS + GriPhyN
- Recognized that data will never be centralized
- Continued with NVO (NSF ITR)
- International Virtual Observatory Alliance
  - Now in 15 countries
  - Total data holdings >300TB
- Core services and standards adopted
- Getting ready for first deployment (mid04)

SkyServer.SDSS.org

- A modern archive
  - Raw Pixel data lives in file servers
  - Catalog data (derived objects) lives in Database
  - Online query to any and all
- Also used for education
  - 150 hours of online Astronomy
  - Implicitly teaches data analysis
- Interesting things
  - Spatial data search
  - Client query interface via Java Applet
  - Query interface via Web, Emacs, Perl SOAP,…
  - Popular -- 1% of Terraserver ☺
  - Cloned by other surveys (a template design)
  - Web services are core of it.



Federation: SkyQuery.Net

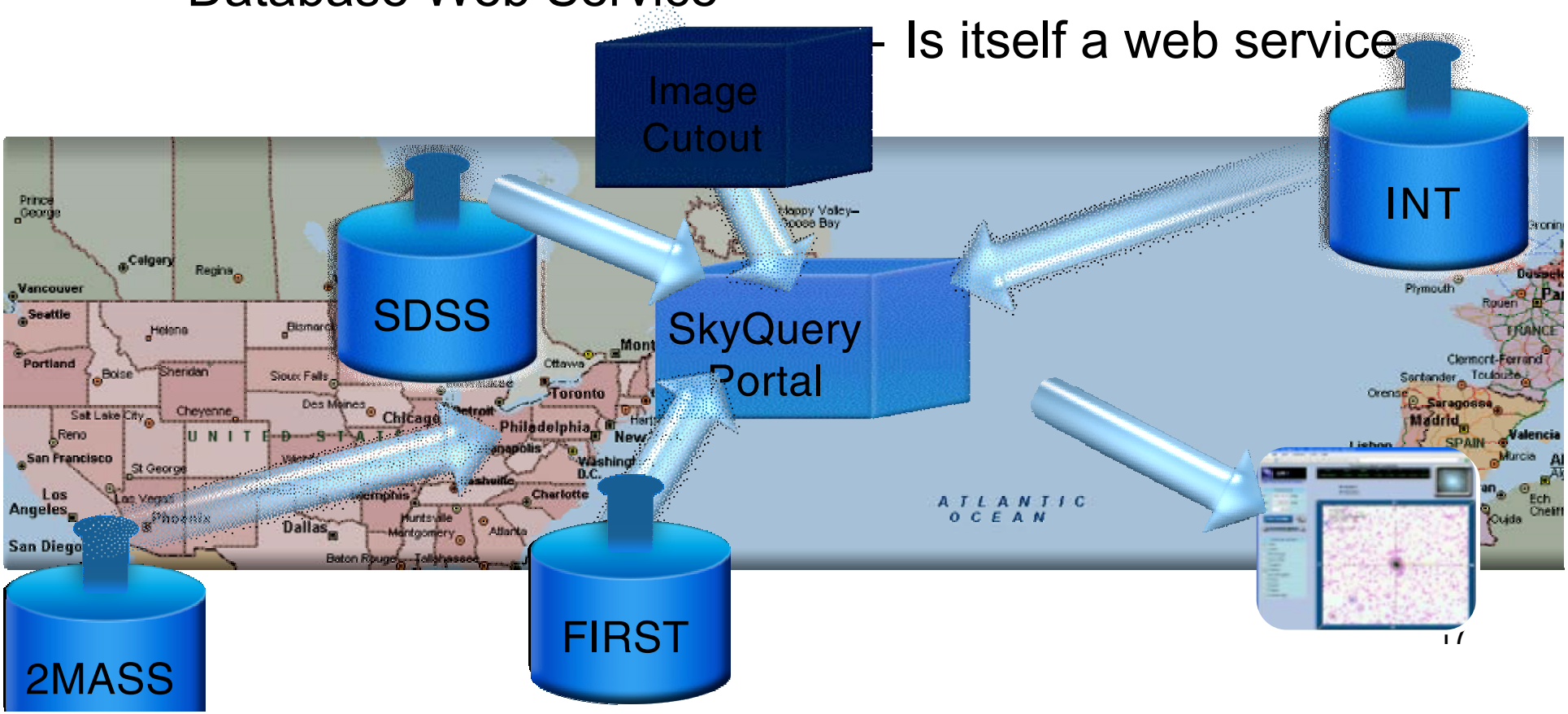


- Combine 4 archives initially
- Now 11 Archives
- Send query to portal,
  portal joins data from archives.
- Problem: want to do multi-step data analys
  (not just single query).
- Solution: Allow personal databases on portal
- Problem: some queries are monsters
- Solution: “batch schedule” on portal server, Deposits answer in personal database.

Extending the SDSS Batch Query System to the National Virtual Observatory Grid

SkyQuery Structure

- Each SkyNode publishes
  - Schema Web Service
  - Database Web Service
- Portal
  - Plans Query (2 phase)
  - Integrates answers
  - Is itself a web service



Interesting Things

- SkyQuery is the most functional Web Service on GriPhyN
- http://skyservice.pha.jhu.edu/develop/vo/adql/
- Now the prototype for an Open Architecture
- Being copied onto Oracle/DB2, Linux, …
- Good test of .NET interop
- Good side-by-side comparison of .NET

Thesis

- Definition:
  - Success == Wide use
- Thesis:
  - Successful Grid App == Content + Applications

oh! forgot momentum
"Objecthy Science"